Комбинаторные оценки вероятности переобучения на основе покрытий множества алгоритмов

А. И. Фрей, И. О. Толстихин

26 августа 2013 г.

Повышение точности оценок обобщающей способности уже более сорока лет остаётся открытой проблемой в теории статистического обучения [1, 2]. Комбинаторный подход впервые позволил получить точные оценки для некоторых модельных частных случаев [3, 5]. Более общие, но менее точные комбинаторные оценки вероятности переобучения, основанные на принципе порождающих и запрещающих множеств, были предложены в [4]. Их применение к семействам конъюнкций пороговых решающих правил позволило улучшить качество логических алгоритмов классификации при решении прикладных задач. Однако применимость этих оценок ограничивалась семействами невысокой мощности, обладающими специальным свойством связности.

Оценки, предлагаемые в данной работе, существенно расширяют границы применимости комбинаторного подхода. Они основаны на покрытии семейства множествами специального вида, для которых точные оценки вероятности переобучения выписываются в явном виде.

Пусть задана генеральная выборка $\mathbb{X} = \left(x_1, \ldots, x_L\right)$ из L объектов. Пусть A — некоторое множество алгоритмов классификации. Произвольный алгоритм $a \in A$, примененный к выборке $\mathbb{X}$, порождает бинарный вектор ошибок $a \equiv \left(I(a, x_i)\right)_{i=1}^{L}$, где $I(a, x_i) \in \{0, 1\}$ — индикатор ошибки алгоритма a на объекте x_i. Для произвольной подвыборки $U \subseteq \mathbb{X}$ *чис-*

ло и *частота* ошибок алгоритма a обозначаются, соответственно, через $n(a,U) = \sum\limits_{x_i \in U} I(a,x_i)$ и $\nu(a,U) = n(a,U)/|U|$.

Пусть $[\mathbb{X}]^\ell$ — множество всех разбиений генеральной выборки $\mathbb{X}$ на обучающую выборку X длины ℓ и контрольную выборку $\bar{X}$ длины $k = L - \ell$. *Методом обучения* называют отображение $\mu\colon [\mathbb{X}]^\ell \to A$, которое произвольной обучающей выборке $X \in [\mathbb{X}]^\ell$ ставит в соответствие некоторый алгоритм $\mu X \in A$. Для произвольного разбиения $\mathbb{X} = X \sqcup \bar{X}$ *переобученностью* алгоритма $a = \mu X$ называют уклонение частот его ошибок на контроле и на обучении $\delta(a,X) = \nu(a,\bar{X}) - \nu(a,X)$.

Следуя [3], будем считать, что на множестве $[\mathbb{X}]^\ell$ введено равномерное распределение вероятностей. Пусть $\varphi\colon [\mathbb{X}]^\ell \to \{0,1\}$ — произвольный предикат на $[\mathbb{X}]^\ell$. Тогда вероятность φ есть $\mathsf{P}\,\varphi \equiv \frac{1}{C_L^\ell} \sum\limits_{X \in [\mathbb{X}]^\ell} \varphi(X)$.

Вероятность переобучения $Q_\varepsilon(\mu)$ равна доле разбиений $X \sqcup \bar{X}$, при которых переобученность $\delta(\mu X, X)$ превышает заданный порог $\varepsilon \in (0,1]$:

$$Q_\varepsilon(\mu) = \mathsf{P}[\delta(\mu X, X) \geqslant \varepsilon]. \tag{1}$$

Здесь и далее логическое выражение в квадратных скобках означает, соответственно, $[\mathit{истина}] = 1$, $[\mathit{ложь}] = 0$.

В данной работе оценки вероятности переобучения основаны на разложении множества алгоритмов A на непересекающиеся подмножества:

$$A = A_1 \sqcup A_2 \sqcup \cdots \sqcup A_t, \tag{2}$$

таком что для каждого подмножества A_i можно в явном виде записать условия, при которых $\mu X \in A_i$.

Гипотеза 1. *Пусть множество алгоритмов A представлено в виде разбиения на непересекающиеся подмножества $A = A_1 \sqcup A_2 \sqcup \cdots \sqcup A_t$. Пусть выборка $\mathbb{X}$ и метод обучения μ таковы, что для всех $i = 1,\ldots,t$ можно указать пару непересекающихся подмножеств $X_i \subset \mathbb{X}$ и $X_i' \subset \mathbb{X}$, удовлетворяющую условию*

$$\mu X \in A_i \Rightarrow (X_i \subset X) \text{ и } (X_i' \subset \bar{X}), \quad \forall X \in [\mathbb{X}]^\ell.$$

Пусть, кроме этого, все алгоритмы $a \in A_i$ не допускают ошибок на X_i и ошибаются на всех объектах из X_i'.

Множество X_i будем называть *порождающим*, а множество X'_i — *запрещающим* для A_i. Гипотеза 1 означает, что результат обучения может принадлежать A_i, только если в обучающей выборке X находятся все порождающие объекты и ни одного запрещающего. Все остальные объекты из $Y_i \equiv \mathbb{X} \backslash X_i \backslash X'_i$ будем называть *нейтральными* для A_i.

Для каждого $i = 1, \ldots, t$ введем обозначения $L_i = L - |X_i| - |X'_i|$, $\ell_i = \ell - |X_i|$, $k_i = k - |X'_i|$. Пусть $Q_\varepsilon(A_i)$ есть верхняя оценка вероятности переобучения для произвольного метода $\mu\colon [Y_i]^{\ell_i} \to A_i$:

$$Q_\varepsilon(A_i) = \frac{1}{C_{L_i}^{\ell_i}} \sum_{Y \in [Y_i]^{\ell_i}} [\max_{a \in A_i} \delta(a, Y) \geqslant \varepsilon], \tag{3}$$

где $[Y_i]^{\ell_i}$ — множество разбиений Y_i на обучающую выборку Y длины ℓ_i и контрольную выборку $\bar{Y}$ длины $k_i = L_i - \ell_i$.

Теорема 1 (Оценка расслоения-сходства). *Пусть выполнена гипотеза 1, а на разбиение $A = A_1 \sqcup A_2 \sqcup \cdots \sqcup A_t$ наложено дополнительное ограничение: внутри каждого кластера A_i все алгоритмы допускают равное число ошибок (обозначаемое через m_i). Тогда*

$$Q_\varepsilon(\mu) \leqslant \sum_{i=1}^{t} P_i \, Q_{\varepsilon_i}(A_i), \tag{4}$$

где $P_i = \frac{C_{L_i}^{\ell_i}}{C_L^\ell}$, $\varepsilon_i = \frac{L_i}{\ell_i k_i} \frac{\ell\, k}{L} \varepsilon + \left(1 - \frac{\ell\, L_i}{L \ell_i}\right) \frac{m_i}{k_i} - \frac{|X'_i|}{k_i}$, $Q_\varepsilon(A_i)$ — вероятность переобучения на множестве нейтральных объектов (3).

Чтобы вычислить оценку (4), нужно в явном виде построить порождающие и запрещающие множества для всех A_i, $i = 1, \ldots, t$. Для этого необходимо фиксировать метод обучения.

Определение 1. *Метод обучения μ называют пессимистической минимизацией эмпирического риска (ПМЭР), если для любой обучающей выборки X выполнено $\mu X \in \operatorname*{Argmax}_{a \in A(X)} n(a, \mathbb{X})$, $A(X) \equiv \operatorname*{Argmin}_{a \in A} n(a, X)$.*

Следуя [5], введем на A отношение частичного порядка: $a \leqslant b$ тогда и только тогда, когда $I(a, x) \leqslant I(b, x), \forall x \in \mathbb{X}$. Определим $a < b$ если $a \leqslant b$ и $a \neq b$. Если $a < b$ и $\rho(a, b) = 1$, то будем говорить, что a *предшествует*

b, и записывать $a \prec b$. Здесь и далее $\rho(a_1, a_2) = \sum\limits_{x \in \mathbb{X}} [a_1(x) \neq a_2(x)]$ — *хэммингово расстояние* между алгоритмами a_1 и a_2.

Для отдельного алгоритма $a \in A$ общий способ построения порождающих и запрещающих множеств дается в [5]:

$$\begin{aligned} X_a &= \{x \in X \colon \exists b \in A \colon a \prec b, I(a,x) < I(b,x)\}, \\ X'_a &= \{x \in X \colon \exists b \in A \colon b < a, I(b,x) < I(a,x)\}. \end{aligned} \tag{5}$$

Лемма 2. *Пусть метод обучения μ является ПМЭР. Определим*

$$X_i = \bigcap_{a \in A_i} X_a, \quad X'_i = \bigcap_{a \in A_i} X'_a, \tag{6}$$

где X_a and X'_a определены в (5). Эти множества X_i и X'_i являются, соответственно, порождающим и запрещающим множествами для кластера A_i в смысле гипотезы 1.

Лемма 2 позволяет вычислить оценку (4), за исключением выражения $Q_{\varepsilon_i}(A_i)$. Чтобы исправить это, расширим каждое подмножество A_i до множества $B_i \supseteq A_i$ и заменим $Q_{\varepsilon_i}(A_i)$ на $Q_{\varepsilon_i}(B_i)$.

Теорема 3. *Пусть $A_i \subseteq B_i$, и все алгоритмы из B_i допускают равное число ошибок на полной выборке. Тогда*

$$Q_\varepsilon(A_i) \leqslant Q_\varepsilon(B_i). \tag{7}$$

Теорема 3 подразумевает, что для вероятности переобучения $Q_\varepsilon(B_i)$ известна точная вычислительно эффективная формула. Следующие два определения дают примеры множеств B_i, для которых это выполнено.

Определение 2. *Пусть a_0 — произвольный алгоритм с m ошибками, $n(a_0, \mathbb{X}) = m$, $r \in \mathbb{N}$, $r \leqslant m$. Множеством $B_r^m(a_0)$ будем называть центральный слой хэммингова шара радиуса r с центром в a_0:*

$$B_r^m(a_0) = \{a \in \{0,1\}^L \colon \rho(a, a_0) \leqslant r \text{ и } n(a, \mathbb{X}) = m\}.$$

Определение 3. *Пусть $\mathbb{X} = X_0 \sqcup X_1 \sqcup X_r$ — разбиение генеральной выборки, $|X_r| = r$, $|X_1| = m$, $\rho \in \mathbb{N}$, $\rho \leqslant r$. Множеством $B_{r,\rho}^m$ будем называть подмножество $\{0,1\}^L$ такое, что*

- $B^m_{r,\rho}$ *содержит все алгоритмы, допускающие ровно* ρ *ошибок на* X_r,
- *ни один алгоритм из* $B^m_{r,\rho}$ *не ошибается на объектах из* X_0,
- *все алгоритмы из* $B^m_{r,\rho}$ *ошибаются на всех объектах из* X_1.

При выводе точных формул вероятности переобучения $Q_\varepsilon(B^m_r(a_0))$ и $Q_\varepsilon(B^m_{r,\rho})$ используется рандомизированный метод обучения и теоретико-групповой подход [6, 7]. Оба рассмотренных выше семейства содержат лишь алгоритмы с равным числом ошибок, поэтому оценки, полученные для рандомизированного метода обучения, справедливы также и для детерминированного метода обучения.

Рандомизированный метод минимизации эмпирического риска выбирает произвольный алгоритм из $A(X)$ случайно и равновероятно [6]. При этом в определение вероятности переобучения (1) приходится вводить дополнительное усреднение по множеству $A(X)$:

$$Q_\varepsilon(A) = \frac{1}{C^\ell_L} \sum_{X \in [\mathbb{X}]^\ell} \frac{1}{|A(X)|} \sum_{a \in A(X)} \big[\delta(a, X) \geqslant \varepsilon\big]. \tag{8}$$

Пусть $S_L = \{\pi \colon \mathbb{X} \to \mathbb{X}\}$ — симметрическая группа из L элементов, действующая на генеральную выборку перестановками объектов. Действие произвольной $\pi \in S_L$ на алгоритм $a \in A$ определено перестановкой координат вектора ошибок: $(\pi a)(x_i) = a(\pi^{-1} x_i)$. Для произвольной выборки $X \in [\mathbb{X}]^\ell$ и множества алгоритмов $A \subset \{0,1\}^L$ действия πX и πA определены следующим образом: $\pi X = \{\pi x \colon x \in X\}$, $\pi A = \{\pi a \colon a \in A\}$.

Определение 4. *Группой симметрий* $\mathrm{Sym}(A)$ *множества алгоритмов* $A \subset \{0,1\}^L$ *называется его стационарная подгруппа:*

$$\mathrm{Sym}(A) = \{\pi \in S_L \colon \pi A = A\}.$$

Пусть $\Omega([\mathbb{X}]^\ell)$ — множество орбит действия группы $\mathrm{Sym}(A)$ на $[\mathbb{X}]^\ell$. Произвольного представителя орбиты $\tau \in \Omega([\mathbb{X}]^\ell)$ обозначим через X_τ.

Теорема 4. *Пусть* A *— множество алгоритмов, и* $\mathrm{Sym}(A)$ *— его группа симметрий. Тогда вероятность переобучения* (8) *записывается в виде*

$$Q_\varepsilon(A) = \sum_{\tau \in \Omega([\mathbb{X}]^\ell)} \frac{|\tau|}{|A(X_\tau)|} \sum_{a \in A(X_\tau)} [\delta(a, X_\tau) \geqslant \varepsilon]. \tag{9}$$

Данная теорема позволяет записать вероятность переобучения $Q_\varepsilon(B_r^m(a_0))$ и $Q_\varepsilon(B_{r,\rho}^m)$ в следующем виде.

Теорема 5. *Вероятность переобучения для $B_r^m(a_0)$ при $r \leqslant 2m$ и $n(a_0, \mathbb{X}) = m$ записывается в виде*

$$Q_\varepsilon(B_r^m(a_0)) = H_L^{\ell,m}\left(s(\varepsilon) + \left\lfloor r/2 \right\rfloor\right) \cdot [m \geqslant \varepsilon k], \tag{10}$$

где $s(\varepsilon) = \frac{\ell}{L}(m - \varepsilon k)$, $H_L^{\ell,m}(s) = \sum\limits_{t=0}^{\lfloor s \rfloor} C_m^t C_{L-m}^{\ell-t} / C_L^\ell$ — функция гипергеометрического распределения [3].

Теорема 6. *Вероятность переобучения для $B_{r,\rho}^m$ записывается в виде*

$$Q_\varepsilon(B_{r,\rho}^m) = \frac{1}{C_L^\ell} \sum_{i=0}^{\min(m,\ell)} \sum_{j=0}^{\min(r,\ell-i)} C_m^i C_r^j C_{L-m-r}^{\ell-i-j} \left[\frac{m+\rho-t}{k} - \frac{t}{\ell} \geqslant \varepsilon\right], \tag{11}$$

где $t = i + \max(0, \rho - r - j)$.

Предлагается следующий способ вычисления комбинаторной оценки расслоения-сходства для произвольного множества алгоритмов A. Множество A разбивается на кластеры (2), для каждого из которых строится порождающее и запрещающее множество (6). Затем каждый кластер A_i вкладывается в объемлющее множество B_i, для которого известна точная формула вероятности переобучения, например, (10) или (11). Итоговая оценка записывается в виде (4).

Численный эксперимент на 11 задачах из репозитория UCI показал, что оценка расслоения-сходства, вычисленная предложенным способом, даёт верхние оценки частоты ошибок на контрольной выборке, завышенные лишь на 5 – 50 % по сравнению с фактической частотой ошибок на контроле. На тех же задачах завышенность оценок расслоения-связности из [5] составляет 17 – 63 %. Наиболее точные из PAC-Bayes оценок, предложенные в недавней работе [8], оказались завышены в 2,5 – 10 раз.

Работа поддержана РФФИ (проект №11-07-00480, №12-07-33099-мол-а-вед) и программой ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения».

Список литературы

[1] Vapnik V. N., Chervonenkis A. Y. (1971) On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2), 264–280.

[2] Boucheron S., Bousquet O., Lugosi G. (2005) Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9(1), 323–375.

[3] *Воронцов К. В.* Точные оценки вероятности переобучения // Доклады РАН, 2009. — Т. 429, № 1. — С. 15–18.

[4] Vorontsov K. V., Ivahnenko A. A. (2011) Tight combinatorial generalization bounds for threshold conjunction rules. *4-th Int'l Conf. on Pattern Recognition and Machine Intelligence (PReMI'11).* Lecture Notes in Computer Science, Springer-Verlag, 66–73.

[5] *Vorontsov K. V.* Exact combinatorial bounds on the probability of overfitting for empirical risk minimization // Pattern Recognition and Image Analysis. — 2010. — Vol. 20, No. 3. — Pp. 269–285.

[6] *Фрей А. И.* Точные оценки вероятности переобучения для симметричных семейств алгоритмов и рандомизированных методов обучения // Pattern Recognition and Image Analysis. — 2010.

[7] *Толстихин И. О.* Вероятность переобучения плотных и разреженных семейств алгоритмов // Междунар. конф. ИОИ-8 — М.: МАКС Пресс, 2010. — С. 83–86.

[8] Jin C., Wang L. (2012) Dimensionality Dependent PAC-Bayes Margin Bound. *In Advances in Neural Information Processing Systems*, 25, 1043–1051.

УДК 519.7:004.855.5

Перевод названия, имени и фамилии авторов

Cover-based combinatorial bounds on probability of overfitting

Alexander Frey, Ilya Tolstikhin

Данные об авторах

1. Фрей Александр Ильич

Тел.: +7(903)175-80-77;

Email: sashafrey@gmail.com;

Московский Физико-Технический Институт (Государственный Университет)

2. Толстихин Илья Олегович

Тел.: +7(916)136-49-78;

Email: iliya.tolstikhin@gmail.com;

Учреждение Российской академии наук Вычислительный центр им. А. А. Дородницына РАН.

Реферат

Комбинаторные оценки вероятности переобучения на основе покрытии множества алгоритмов

А. И. Фрей, И. О. Толстихин

Предлагается новая комбинаторная оценка вероятности переобучения, учитывающая сходство алгоритмов. Оценка основана на разложении множества алгоритмов на непересекающиеся подмножества (кластеры). Каждый кластер пополняется алгоритмами до объемлющего множества алгоритмов с известной точной оценкой вероятности переобучения. Итоговая оценка учитывает сходство алгоритмов внутри каждого кластера, и расслоение алгоритмов по числу ошибок между разными кластерами. Для вывода вероятности переобучения объемлющих множеств предлагается теоретико-групповой подход, основанный на учете симметрий множества алгоритмов. Приводятся два частных примера симметричных семейств алгоритмов, для которых теоретико-групповой подход позволяет выписать точную вычислительно-эффективную оценку вероятности переобучения.